Document Clustering Based On Text Mining K-Means Algorithm Using Euclidean Distance Similarity

Article *in* Journal of Advanced Research in Dynamical and Control Systems · April 2018

CITATIONS
28

READS
6,517

4 authors, including:

Laxmi Lydia
Vignan's Institute of information Technology
195 PUBLICATIONS **2,118** CITATIONS
SEE PROFILE

P. Govindasamy
Anna University, Chennai
5 PUBLICATIONS **63** CITATIONS
SEE PROFILE

S.K. Lakshmanaprabu
Renault Nissan Technology and Business Center India Pvt Ltd
65 PUBLICATIONS **3,925** CITATIONS
SEE PROFILE

Document Clustering Based On Text Mining K-Means Algorithm Using Euclidean Distance Similarity

[1] *E. Laxmi Lydia,*[2]*P.Govindaswamy,*[3] *SK.Lakshmanaprabu,* [4]*D.Ramya*

[1]Associate Professor, Department of CSE, Vignan's Institute of Information Technology, Andhra Pradesh, India.
[1]Research Scholar, Department of Industrial Engineering, Anna University, CEG Campus, Chennai, India

[3]Research Scholar Electronics and Instrumentation Engineering, B S Abdur Rahman Crescent Institute of Science and Technology, Chennai, India

[4]Junior Research Fellowship(JRF),Department of CSE, Vignan's Institute of Information Technology, Andhra Pradesh, India.

Abstract- Data mining a specific area named text mining is used to classify the huge semi structured data needs proper clustering. Maximum text documents involves fast retrieval of information, arrangement of documents, exploring of information from the documents .Declaration of text input data and classification of the documents is a complex process. The main objective of this paper is to produce a specific open source to class the clusters of identical documents in the interrelated folders and to lower the complexity of locating each document. Algorithms considered are challenges for open research responsibilities. This paper describes the document clustering process based on the clustering techniques, partitioning clustering using K-means and also calculates the centroid similarity and cluster similarity.

Keywords : Document Clustering, TF-IDF, Clustering techniques, K-means

Introduction

A lot of text document sources are suitable for manipulating computerized computations. Text mining deals with unorganized , semi-structured datasets. Techniques involved in text mining initiates with the selection of text documents.

The pre-processing approaches unclutters and formats the input data, also accordingly the extraction of useful feature information from the documents is done. Later the techniques in text mining such as clustering, classification algorithms are designed to position the documents. Various classification approaches for the process of discover a set of models or functions that illustrate and categorize data classes or consideration for the objective of visualizing the class objects whose class label is not specified and clustering techniques figure out data objects without examine the known class model. Following figure1 describes the different techniques that involve in clustering. Among all clustering techniques K-means provides the best efficiency.

ISSN 1943-023X
Received: 5 Mar 2018/Accepted: 10 Apr 2018

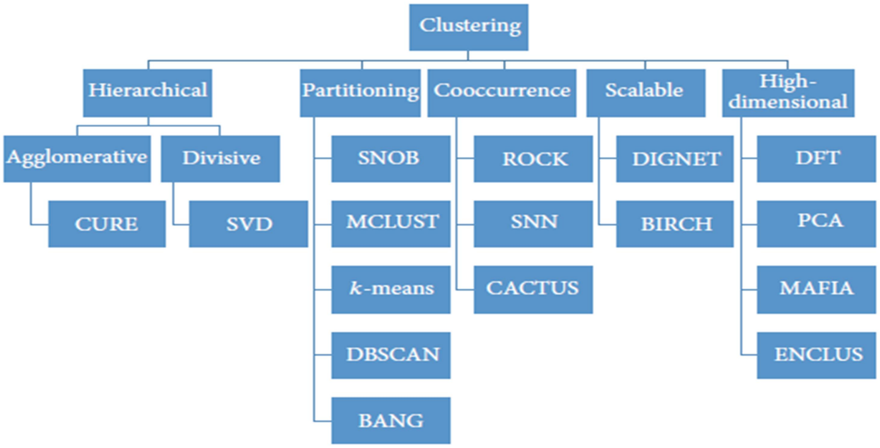

Figure 1 : Structure of Clustering techniques

Term Document Matrix

After pre-processing, a table is maintained based on the terms in the documents depending on occurrences of terms in each document calculating their frequencies. Importance of the term in each document is calculated based on the weight functions and the entire collection of the document. Every document comprises of particular words, this table creates a high dimensional and sparse features which brings a tremendous noise to the text clustering and makes it hard for clustering algorithms to appropriately cluster identical documents.

TF-IDF(Term Frequency- Inverse Document Frequency)

TF-IDF weight is generally used in text mining for information retrieval. Here the weight is identified to calculate how essential is the word to a document in a set. The weight increases correspondingly to the number of times a word occur in the document.

TF(Term Frequency) :

Term Frequency describes the occurrence of the term that is most regular in the document. Each document length is different, there are changes that a term may appear more number of times in bigger documents than in smaller ones. Therefore the term frequency is given by the following calculations:

$$TF(t) = \frac{Number\ of\ times\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}.$$

IDF(Inverse Document Frequency) :

Inverse Document Frequency describes the importance of the term. when calculating term frequency , the terms acknowledge the equal importance. Some terms may appear more frequently but have less importance. Therefore we need to calculate the weights of the frequent terms and note the rare terms by evaluating . The Inverse document frequency is given by the following calculations:

$$IDF(t) = \log \frac{Total\ number\ of\ document}{Number\ of\ documents\ with\ term\ t\ in\ it}$$

Literature Survey

Document Clustering is extensively used text mining ranging the capability with the growth in possibility of available text data. Text document clustering is applied to certainly to a group of document that associate to the same topic to provide users peruse of improved results [3].

Received: 5 Mar 2018/Accepted: 10 Apr 2018

Experimental information confirm that the prosperity from document clustering[1]. Document clustering is consistently been used as a mechanism to enhance the achievement of improvement and operate spacious data. Currently clustering has been advanced for reading a collection of documents .

Balabantaray et al., [2] correlates the K-means clustering with K-mediods clustering. K-means was performed based on Euclidean distance and Manhattan distance measures in WEKA. Lastly, it was examined that K-means produce improved outcome than K-Mediods.

Greene Derek et al.,[4] popularized text clustering with groundwork advanced unsupervised text mining works. Jain et al.,[5] deliberates regarding pre-processing of documents, operations of text clustering. and also with their pros and cons of text clustering along with some key approaches that finalizes the algorithms which allow not to perform overlapping of clusters.

Jajoo et al.,[6] explains how to progress the performance and accuracy of clustering in documents. Techniques like partitioning clustering applied in document clustering to execute more desirable performance results than standard clustering algorithms. This decreases the noise in data and enlights the quality of clusters.

Khadhim et al., [7] Implemented TF-IDF and SVD dimensionality reduction techniques and implemented the reduction techniques and given the performance for text clustering in documents..

Liu Tao et al.,[9] proven that feature selection approaches can increase the efficiency and accuracy of algorithms in clustering. The term condition is favorable than DF and Entrophy based.

Mugunthadevi et al.,[11] worked and studied on various feature selection approaches along with their pro and cons. Lastly ended by stating that the feature selection holds on the field by giving better results facing new challenges in text mining.

Tang bin et al.,[12] handles a transformation mechanism considerably to diminish the computational cost combined with the finest transformation approaches like Independent component Analysis (ICA) and Latent Semantic Indexing (LSI) defending the clustering accomplishment. .

Zhoa et al.,[13] recommended a new mechanism that initiates the cloud model theory to feature selections in building up clustering documents. Practical outcomes with K-means algorithm shown in the field has a significant enhancement in circumstances of accuracy of text clustering.

Methodology

Procedure involved for Document Clustering:

Pre-processing of text: Pre-processing comprise of removing of unwanted noise in the textual data using Stop words algorithm for each document.
Feature Generation: Features are generated by using the root of the words on applying Stemming algorithms.
Feature Extraction: After generating of the words , each stem word is calculated for their weights using TF-IDF (Term Frequency- Inverse Document Frequency). After assigning the weights , decrease the number of features and select features only that have maximum weights in the document.
Clustering: After getting the term document matrix with weights apply partitioning clustering algorithm to group the documents into K clusters.
Result Evaluation: Lastly calculations and analysis are carried out by clustering method. Below figure2 gives the flow chart diagram for entire process evolved in Document clustering.

ISSN 1943-023X
Received: 5 Mar 2018/Accepted: 10 Apr 2018

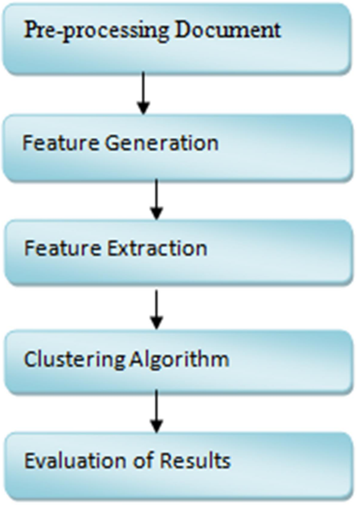

Figure 2: General Flow chart steps for Document clustering

The term Clustering itself defines grouping of similar objects in one cluster. Implementing of clustering algorithms also maintain same principle by dividing the data based on the features that are extracted from the document. the algorithm considered in this project is Partitioning algorithm. Clusters involve unsupervised learning. Partitioning algorithms implement clusters discover clusters either by repetitive rearranging of points among subgroups or identifying areas that are effected mostly with the data. When the features are more sparsely increases with respect to high dimensions. Best results are evaluated when use of K-means when the data considered is high dimensional clustering using partitioning clustering.

Procedure For K-Means Algorithm

K-means algorithm is most efficient algorithm that is targeted to the unsupervised learning. Here algorithms directs to partition a set of objects related to the features in K clusters. The partitioning algorithm the documents need to be clustered where each cluster need to specify the centroid. the centroid is calculated by the average mean of the objects that exist in the cluster. Assigning centroid to the each cluster is the main key step for the algorithm. Assigning of centroid in clusters is based on the similarity function that is carried out by Euclidean distance to all the objects that are existing in the cluster. the following is the algorithm steps for the K-means partitioning algorithm with respect to the input and the result outcome.

```
Input : a dataset containing n-dimensional term document matrix as input and assume number of clusters(K)
Output: Set of K clusters.

Procedure :
Step 1: Consider K numbers of clusters.
Step 2: Consider $C_k$ centroid randomly to initialize centers of the each cluster based on the mean.
Step 3: Repeat
        3.1 Every cluster center is assigned with closed object by calculating distance using
        Euclidean similarity measure.
        3.2 Calculate the new centroid points on each step by mean points.
Step 4 : Until
        4.1 Calculate a until cluster centers remain constantly without any change
        4.2 Also consider that no object changes its clusters.
```

Architectural Design

ISSN 1943-023X
Received: 5 Mar 2018/Accepted: 10 Apr 2018

The design that is implemented in this project is to group similar documents in clusters. To obtain the procedure , the documents initially needs to be pre-processed and then term document matrix will be generated. The term document matrix is represented by the collection of the terms that are calculated based on the weights of the words in the document. Finally these clusters identify the documents in order

Document Clustering Procedure

1. Read all the input documents.
2. Determine all the unique terms from the considered input documents.
3. Generation of Input vectors using TF-IDF values in n-dimensional term-document matrix.
4. Hire the document vector to K-means clustering algorithm.
5. Selection of similarity measure for generating similarity matrix using Euclidean distances.
6. Observe the clusters and measure performance.
7. Calculate centroid similarity and cluster similarity for even more effective results.

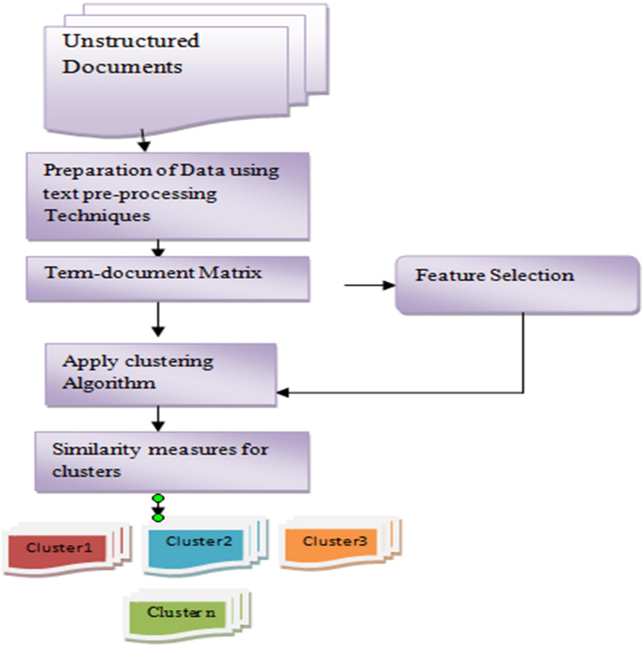
Unstructured Documents
Preparation of Data using text pre-processing Techniques
Term-document Matrix
Feature Selection
Apply clustering Algorithm
Similarity measures for clusters
Cluster1
Cluster2
Cluster3
Cluster n

Figure 3: Flow chart for Document Clustering

Result Analysis

Here for the process of clustering, we have considered three documents, and assumed three clusters basing on the K-Means algorithm implementing Euclidean Distance similarity measure.

The input data is retrieved from 20Newsgroup and particularly the input values are given by the TF-IDF values of considered documents. For the process of text mining , after the generation of term weights , distance is calculated and clustering algorithm is performed. Here we considered terms based similarity on each document and resulted clusters as output.

ISSN 1943-023X
Received: 5 Mar 2018/Accepted: 10 Apr 2018

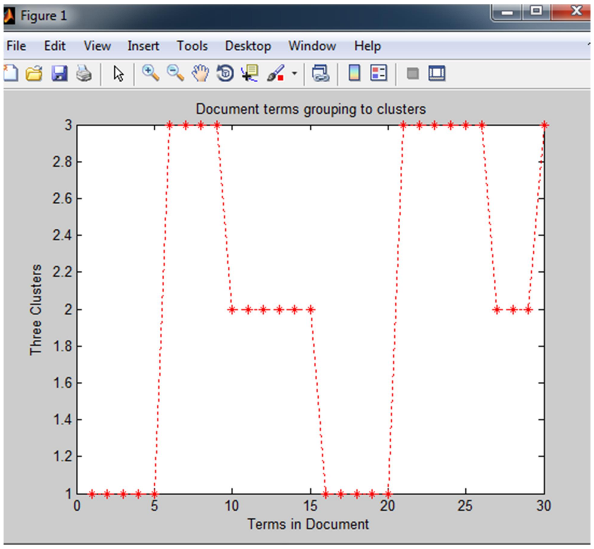

Figure 4: Graph plotting three document terms in clusters

The above graph figure 4 describes the results based on the three document terms. Here all the three documents (doc1, doc2, doc3) terms are evaluated using term similarity distance measure performing Euclidean distance and the obtained distances based on the centroids documents are grouped. The process is repeated till the documents lies on the same cluster due to Centroids assumptions and mean calculations. Finally after calculating all the terms in the document, we group similar documents in one cluster. The cluster may contain more number of documents that are identical.

Conclusion

The main aim of this current study is to use text mining techniques on unstructured data in documents. This paper works on the detailed working process involved in the K-means algorithm and in document pre-processing. This paper was proposed for finding similarities among the document terms based on the TF-IDF weights, where similar documents are grouped in one cluster.

FUTURE SCOPE

Data need to be read and write that need to be in TB, PB and ZB for fast and efficient mechanisms. Mining algorithms need to handle uncertain data in different applicational areas. The most open issues are data checking, parallel programming models for the security of data, privacy, the data sharing mechanisms, the growth of data size. These must be enhanced easily when we have fast efficient clustering algorithms.

References

[1] Andrews, Nicholas O., and Edward A. Fox " Recent development in document clustering." (2007).
[2] Balabantaray , Rakesh Chandra, Chandrali Sarma, and Monica Jha. " Document Clustering using K-Means and K-Medoids." arXiv preprint arXiv:1502.07938 (2015)
[3] Gao, Jing, and Jun Zhang. "Clustered SVD strategies in latent semantic indexing."Information processing & management 41.5 (2005): 1051-1063.
[4] Greene, Dereck. A State-of-the-art Toolkit for Document Clustering. Diss. Trinity College,2007.
[5] Jain, Yogesh, and Amit Kumar Nandanwar. "A Theretical Study of Text Document Clustering."
[6] Jajoo, Pankaj. Document clustering. Diss.Indian Institute of Technology Kharagpur,2008.
[7] Kadhim, Ammar Ismael, Yu-N. Cheah, and Nurrul Hashimah Ahamed. "Text Document Pre-processing and Dimension Reduction Techniques for Text Document Clustering." Artificial Intelligence with Applications in Engineering and Technology (ICAIET),2014 4th International Conference on. IEEE,2014.

ISSN 1943-023X
Received: 5 Mar 2018/Accepted: 10 Apr 2018

[8] U.S. Patki, Dr. P.G. Khot, A Literature Review on Text Document Clustering Algorithms used in Text Mining Journal of Engineering Computers & Applied Sciences (JECAS) ISSN No: 2319-5606 Volume 6, No.10, October 2017

[9] Liu, Tao, et al. "An evaluation on feature selection for text Clustering." Icml. Vol. 3.2003.

[10] Ahmed Elragal, Moutaz Haddara, Big Data Analytics: A Text Mining- Based Literature Analysis, Department of Computer Science, Electrical and Space Engineering.

[11] Mugunthadevi,K., et al. " Survey on feature selection in document clustering." International Journal on Computer Science and Engineering 3.3 (2011): 1240-1241.

[12] Tang, Bin, et al. " Comparing and combining dimension reduction techniques for efficient text clustering . " Proceedings of SIAM International Workshop on Feature Selection for Data Mining. 2005.

[13] Zhao, Junmin, Kai Zhang, and Jian Wan. "Research of feature selection for text clustering based on cloud model. " Journal of Software 8.12 (2013): 3246-3252.

[14] Yogapreethi.N,Maheswari.S, A Review On Text Mining In Data Mining International journal on soft computing (IJSC) Vol. 7,No. 2/3, August 2016.

[15] Feng Chen,1,2 Pan Deng,1 Jiafu Wan,3 Daqiang Zhang,4 Athanasias V . Vasilakos,5 and Xiaohui Rong6, Data Mining for the Internet of Things: Literature Review and Challenges International Journal of Distributed Sensor Networks Volume 2015, Article ID 431047

[16] T.V. Rajinikanth 1 and G. Suresh Reddy2 , A Soft Similarity Measure For K-Means Based High Dimensional Document Clustering, IADIS International Journal on Computer Science and Information Systems Vol. 12, No. 1, pp. 88-108ISSN:1646-3692.

[17] A. Sudha Ramkumar,Dr. B. Poorna, Text Document Clustering Using Dimension Reduction Technique, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 7 (2016) pp 4770-4774.

[18] R. Balamurugan 1, Dr. S. Pushpa2 1Research Scholar, 2Professor, Computer Science and Engineering, St. Peter's University, Chennai (India),A Review On Various Text Mining Techniques And Algorithms , 2nd International conference on Recent Innovations in Science, Engineering and Management.

[19] Monika Gupta, 2Kanwal Garg , A Review on Document Clustering, International Journal of Advanced Research in Computer Science and Software Engineering, Volume6,Issue 5, May 2016.

Funding Acknowledgment

This work is financially supported by the Department of Science and Technology(DST), Science and Engineering Research Board(SERB) under the scheme of ECR. we thank DST_SERB for the financial support to carry the research work.

ISSN 1943-023X

Received: 5 Mar 2018/Accepted: 10 Apr 2018